DOCUMENTATION OF DECISION-AIDING SOFTWARE:
SCORING RULE USERS MANUAL

DECISIONS AND DESIGNS INC.

Dorothy M. Amey
Phillip H. Feuerwerger
Roy M. Gulick

July 1979

N00014-79-C-0069

ADVANCED DECISION TECHNOLOGY PROGRAM

CYBERNETICS TECHNOLOGY OFFICE
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
Office of Naval Research • Engineering Psychology Programs

(1)

DOCUMENTATION OF DECISION-AIDING SOFTWARE:

SCORING RULE USERS MANUAL

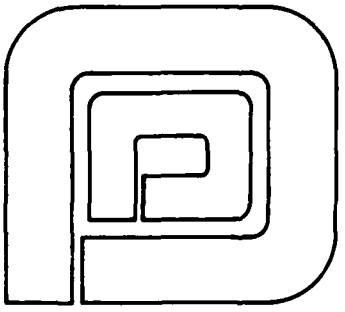
by

Dorothy M. Amey, Phillip H. Feuerwerger, and Roy M. Gulick

Sponsored by

Defense Advanced Research Projects Agency
ARPA Order 3469

July 1979



DECISIONS and DESIGNS, INC.
Suite 600, 8400 Westpark Drive
P.O.Box 907
McLean, Virginia 22101
(703) 821-2828

CONTENTS

FIGURES

SCORING RULE USERS MANUAL

1.0 INTRODUCTION

1.1 Purpose of the Users Manual

The purpose of this manual is to provide the users of Scoring Rule with the background material and the detailed instructions necessary to use and interpret the various functions that the software provides. The manual also presents the technical concepts inherent in the Scoring Rule approach and includes a step-by-step example.

Because the manual must serve users both skilled and unskilled in the use of the methodology, it is prepared in a modular fashion. Thus, whereas the initial sections provide detailed, heavily elaborated information for the naive user, the last section is direct and unelaborated for those users knowledgeable in the approach.

1.2 References

1.2.1 Brown, R.; Kahr, A.; Peterson, C. Decision Analysis for the Manager. New York: Holt, Rinehart, and Winston, 1974.

1.2.2 Chinnis, J. O., Jr.; Feuerwerger, P. H.; Kelly, C. W., III. A Procedure for Evaluating the Subjective Probability Assessment Test. Technical Report PR 78-19-74. McLean, Virginia: Decisions and Designs, Inc., September 1977.

1.2.3 Amey, Dorothy M.; Feuerwerger, Phillip H.; Gulick, Roy M. Documentation of Decision-Aiding Software: Scoring Rule Functional Description. McLean, Virginia: Decisions and Designs, Inc., July 1979.

1.2.4 Amey, Dorothy M.; Feuerwerger, Phillip H.; Gulick, Roy M. Documentation of Decision-Aiding Software: Scoring Rule Systems Specification. McLean, Virginia: Decisions and Designs, Inc., July 1979.

1.2.5 Barclay, Scott; Brown, Rex V.; Kelly, Clinton W. III; Peterson, Cameron R.; Phillips, Lawrence, D; Selvidge, Judith. Handbook for Decision Analysis. McLean, Virginia: Decisions and Designs, Inc., September 1973.

1.3 Terms and Abbreviations

1.3.1 Scoring Rule - Scoring Rule, the name of the system, is a short description of the function performed by the software, reflecting the system's method for testing, scoring, and training probability assessors.

1.3.2 SCORE - SCORE, an abbreviation for Scoring Rule, is used throughout this report to refer to the system.

1.3.3 SPAT - SPAT is an abbreviation for Subjective Probability Assessment Test, the type of test administered to the individuals being scored and trained by using the SCORE system.

1.3.4 Terms - Standard mathematical notations and decision-analytic terminology are used throughout this Users Manual. Chapter 31 of reference 1.2.1 provides additional background and insight into the basic concepts underlying the procedures implemented by SCORE, as do references 1.2.2 and 1.2.5.

2.0 SYSTEM SUMMARY

2.1 Background

The quality of a decision-making process can be no better than the quality of the analysis of the information bearing on the situation. Conclusive evidence gathered during numerous psychological experiments and clinical observations of working intelligence analysts and their products demonstrates that humans often perform poorly when they assess and communicate to others the relative likelihoods of future events. The impact of that poor performance can be of critical significance to the decision maker.

Suboptimal human performance in assessing and communicating information stems from several factors: the use of qualitative language to express the degree of certainty, which is an inherently quantitative assessment problem; overreliance on intuitive assessment strategies; failure to employ inference processes; deficient probabilistic reasoning; and a tendency to circumvent or ignore formal statistical procedures. All of those factors, together with fundamental cognitive limitations, lead to deficiencies that are observable, measurable, consistent, and predictable; they are biases. Biases in processing information are exhibited by virtually all humans, specifically including experienced forecasters, planners, and intelligence analysts.

The Scoring Rule system described in this manual is designed to discover and display the assessment biases of individual assessors. The goal is to promote discovery and increase the user's awareness of the bias, to provide insight, and to promote improved probability assessments and therefore improved decision making. SCORE is an automated probability assessment testing, scoring, and training system.

The type of test administered by SCORE is also known as a subjective probability assessment test (SPAT).

The overall goal of SCORE is to improve the calibration of probability assessors so as to ensure that their expressed probability assessments are totally consistent with their considered beliefs. Achievement of that goal will facilitate the decision maker's making a decision choice that is consistent with the true beliefs about the likelihood of those future events that will affect the final decision outcome.

2.2 The Communication of Uncertainty

Intelligence analysts and forecasters must assess accurately and communicate effectively with the decision maker to ensure that uncertainty is properly accounted for in the decision-making process. With regard to the communication issue, there is considerable evidence indicating that the use of qualitative terms such as "likely," "probable," and "could" for conveying one's degree of belief about the occurrence of some future event inevitably leads to miscommunication and hence misunderstanding and misperception. For example, consider an experiment conducted with seasoned professional intelligence analysts. The experiment involved an actual intelligence estimate and its conclusion that "... the truce now in effect could continue for six months." Experienced intelligence analysts were asked to read the estimate and then state, based on a probability scale, what they thought the author was trying to convey; that is, to state how likely it was, in the author's mind, that the truce would remain in effect for six months. Several analysts converted the author's qualifier "could" to a likelihood of about 20%. Other analysts placed the likelihood near 50%. Still others perceived that the author felt that a continuation of the truce was almost inevitable--close to 100%. Interestingly and unexpectedly, it turned out that the

estimate had been written by two coauthors. When questioned at different times about the use of the word "could," one author said it was equivalent to a 30% probability of occurrence; the other said 80%.

That experiment and many similar experiments have demonstrated conclusively that very little communication takes place when analysts use qualitative phrases. "Rain is likely" is much more likely to be misperceived than "the chance of rain is 80%."

2.3 Probability

Those whose line of work involves assessing future events should realize that there is but one standard measure for expressing a degree of certainty. That standard measure is probability.

A probability is a number between 0 and 1, inclusive, that represents the extent to which a rational and well-informed individual assessor believes that a future event will occur. It represents a state of mind. The assessor's knowledge may stem from many different sources of information, but the resulting state of mind must ultimately be made explicit and communicated in the form of a probability.

A probability assessment of 1 (100%) indicates that, based on observation, pertinent information, relevant experience, background, and knowledge the assessor is absolutely certain that the event in question will occur. Similarly an assessment of 0 (0%) indicates that the assessor is absolutely certain that the event will not occur. A probability assessment of .5 (50%) indicates that the assessor has no more reason to believe that the event will occur than it won't. If the assessor's state of knowledge is such that the occurrence of the event appears three times more likely

than its failure to occur, then a probability assessment of .75 (75%) would be appropriate. The assessment should communicate the state of the assessor's own knowledge and information base, and no more. SCORE administers a testing procedure that enables probability assessors to discover how well their probability assessments correspond with reality. That is, SCORE is concerned with the problem of calibration bias--ascertaining whether an assessor is too conservative or too confident in assessing probabilities.

2.4 The Testing Procedure

SCORE administers to the user, seated at an interactive terminal, a test consisting of a series of sequential questions. Two alternative answers are displayed simultaneously with each question. In each case, one of the two answers is correct, and the other is incorrect, their order being random. Figure 2-1 shows a typical display presented to the user.

WHICH CITY IS FARTHER WEST:

1. *Reno*
2. *Los Angeles*

Figure 2-1
A SAMPLE DISPLAY

The user must respond to each question as it is presented, citing not only the answer the user believes is correct, but also the user's degree of certainty that the cited answer is indeed the correct one. The degree of certainty is expressed as a probability. For example, the

response to the question in Figure 2-1 might be: 2 .8, which indicates that the user is 80% certain that Los Angeles is west of Reno. A response of 1 1 would indicate that the user is absolutely certain that Reno lies west of Los Angeles.

Since SCORE presents only two answers for consideration, the allowable range for the probability of the answer being correct extends from .5 (completely uncertain as to the correct answer) to 1.0 (absolutely certain). The reason is that a probability of less than .5 would be inconsistent because it would imply that the user believes that the other alternative answer is more likely to be correct. For example, a user's response of 1 .3 (30% certain that the answer is Reno) is logically equivalent to the statement that the user really believes that Los Angeles is the correct answer and not Reno as the user stated.

SCORE incorporates a scoring procedure known as a proper scoring rule. A proper scoring rule is designed to reduce guessing (as discussed in Chapter 30 of reference 1.2.2) by ensuring that the only assessment strategy that will pay off in the long run is a strategy of telling the truth. To implement the proper scoring rule, SCORE displays, following each user response, a win/lose point score that is based on the user's expressed degree of certainty. The higher the expressed probability, the more the user will lose if the cited answer proves to be incorrect. For example, should the user respond: 2 .8 to the Reno/Los Angeles question, SCORE would then display: WIN 21.0 LOSE 39.0, indicating the level of the risk involved. The user should mentally picture a decision model as shown in Figure 2-2.

The rational user should carefully consider the equivalent lottery shown in the figure and mentally reexamine the knowledge base that lead to the answer. If the user is

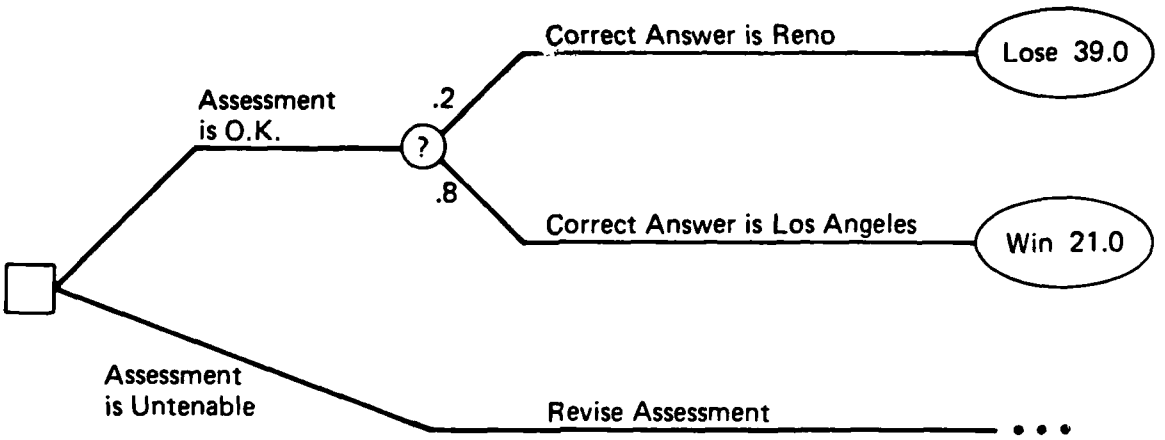


Figure 2-2
EQUIVALENT LOTTERY

overly cautious about accepting the lottery as shown, then the probability assessment should be revised accordingly. SCORE permits the user to revise the assessed probability prior to its revealing the correct answer to the question.

In the case at hand, if the assessment were not revised, SCORE would respond: WRONG YOU LOSE 39.0 POINTS. (Surprisingly, Reno lies west of Los Angeles.)

At the conclusion of the testing session, consisting of perhaps 100 questions, SCORE computes and displays two results, as described later in this manual.

2.5 Probability Assessments

Good probability assessments should reflect three characteristics: they should be rational, coherent, and veridical.

Assessments are rational only to the extent that they are based on observation, high-quality information, experience, and reflection. It is irrational for an assessor to assign a high probability to an event characterized as rare by the assessor's own knowledge base and experience.

Assessments are coherent to the extent to which the assessed probabilities obey the laws of probability theory. An excellent treatment of coherence in probability assessment appears in reference 1.2.1.

Assessments are veridical to the extent that they correspond to reality. A decision maker would soon lose faith in probability assessments if it turned out that in the long run the events that occurred had been associated with low probabilities and those that did not occur had been associated with high probabilities.

The ideal in probability assessments is clairvoyance, that is, the assignment of a probability of 100% to the events that eventually occur and 0% to the events that eventually do not occur. But clairvoyance is, of course, not possible for most important problems. Unfortunately, human beings are not blessed with omniscient powers, and future events are only partially accessible to their foresight.

Although the ideal of clairvoyance is not an obtainable standard, nevertheless it is a standard against which actual probability assessments can be evaluated. Scoring rules, as discussed in Section 2.3, have been developed for the purpose of measuring the veridicality of probability assessments. SCORE uses such a scoring rule and specifically addresses the issue of veridicality.

Veridicality has two components: calibration and resolution. The following example illustrates the difference between those two components.[1]

Intelligence analysts in a Washington Intelligence Agency made weekly forecasts of many different kinds of events such as: whether a military coup would occur within a particular time interval, whether a reconnaissance plane would be shot down, or whether an arms shipment would be made to a particular country within a specified time interval. In each case, it was possible to determine, sometime after the forecast, whether or not the event in question occurred, that is, whether or not the statement for which the probability was assessed turned out to be true.

The probability assessments were evaluated in the following manner. First, the assessments were categorized into intervals of common probabilities. Thus, all assessments near 70% were placed into one category, assessments near 40% were placed into another category, assessments near 10% into yet another category, and so on for all different probability assessments that were used by the analysts. The goal of the analysis was to calculate the percentage of true statements (the hit rate) that fell into each category. In an effort to obtain stable hit rates, the categories each contained approximately 100 different probability assessments.

The percentage of true statements (the hit rate) was calculated within each category by dividing the number of true statements by the total number of statements that were

[1]The remainder of this section is based on material that originally appeared in reference 1.2.5, Barclay, Scott, et al., Handbook for Decision Analysis (Chapter 8, "A Scoring Rule for Probability Assessment").

placed in that category. Figure 2-3 is a calibration diagram that displays the results of this analysis. The vertical axis refers to the average assessed probability for each category. The horizontal axis refers to the corresponding hit rate in each category.

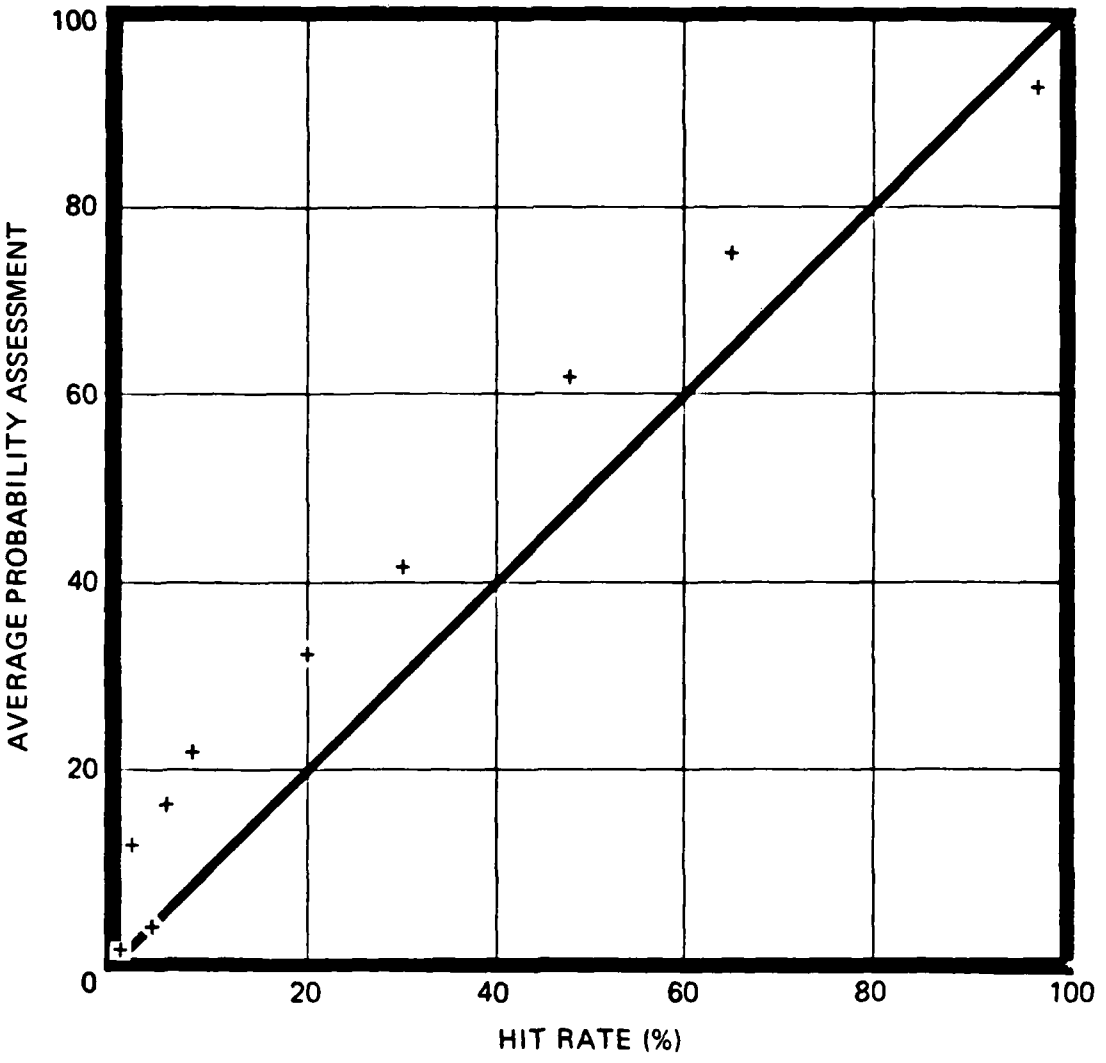


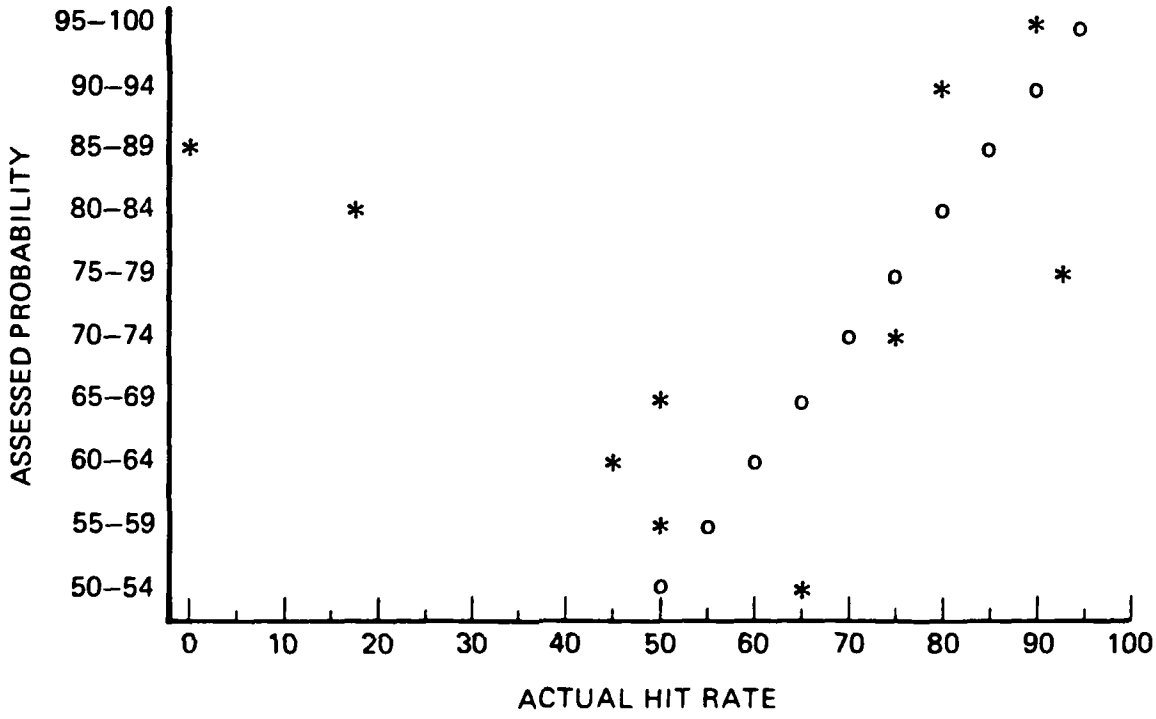
Figure 2-3
CALIBRATION DIAGRAM

The results in Figure 2-3 seem to indicate good performance on the part of the analysts. As the average assessed probability increases, there is an accompanying increase in the percentage of correct statements. An exact measure of performance can be obtained by using a scoring rule to calculate the average veridical error score for all assessments that contributed to each data point in the figure.

The average veridical error for all of the assessments contributing to each data point can be broken into two errors: a calibration error and a resolution error. If an analyst is well calibrated, the data points will all be very close to the diagonal line, as they are in Figure 2-3, where the calibration error of the analysts is very small. A poorly calibrated analyst may produce data points similar to those shown in Figure 2-4, in which the calibration error of the analyst is very large.

The resolution error has to do with the analyst's departure from assessing probabilities of either 100% or 0%; that is, the departure from assessing a future event as either certain or uncertain. Since the event will in fact either occur or not, an assessed probability other than 100% or 0% must contain resolution error. Resolution error is related to the user's level of knowledge of the uncertainty at hand.

At the conclusion of the testing process, SCORE computes and displays a calibration diagram of the type shown in Figure 2-4. This diagram reflects the calibration error. In addition, SCORE computes an overall performance measure. The user's performance is compared with the performance of a perfectly calibrated analyst. Resolution error is ignored.



o: OPTIMAL PERFORMANCE
*: ACTUAL PERFORMANCE

Figure 2-4
CALIBRATION DIAGRAM

The resolution error is removed from consideration because such error is strongly dependent on the analyst's knowledge of the subject matter that comprises the questions. The more the analyst knows about the subject area, the more the assessed probabilities will gravitate toward the end points of the probability scale: 0% and 100%. The resolution error reflects the degree of the analyst's departure from the end points, a departure that is both necessary and desirable for measuring calibration.

SCORE expresses the user's overall performance measure as a percentage of the maximum obtainable calibration score. This result is displayed to the user at the conclusion of the test.

Resolution errors can be reduced primarily by improving the analytic process, either by making more information available or by providing a better means for processing the available information. Calibration errors, on the other hand, can be reduced by improving the process by which probability assessors assign probabilities to reflect their degree of knowledge. A better assignment of probabilities requires an intuitive appreciation for the meaning of a quantitative probability scale. The Scoring Rule test administered by SCORE is intended to improve the user's intuitive understanding of the probability scale, i.e., to reduce calibration error when assigning probabilities.

2.6 SPAT

The SPAT test is a training instrument that is based upon the following rationale. The intention is to provide experience in assessing probabilities for statements or events about which the user had only partial knowledge.

After the probability is assessed, the user learns which event is true and an appropriate error score is assigned by a scoring rule. The user should find, as the test progresses, that a cautious labeling strategy will lead to a relatively poor score. If the user assigns highly uncertain probabilities (near 50%) when he or she is really quite sure of the answer, then the calibration component of the error will be much too high because of that caution. If, on the other hand, a user is too risky, assigning extreme probabilities (near 100% or 0%) when the user is relatively uncertain of the answer, the resolution component of the error score will be too great because of that degree of risk.

The assignment of probabilities that appropriately reflect the user's knowledge about whether or not an event is true should result in a minimum error score that is attributable to the calibration process. Accordingly, as the test unfolds, it is recommended that the user experiment with different calibration strategies. This type of role-playing when taking the test should maximize effectiveness in calibrating probability assessments and improve the ability to make the assessed probability reflect the average hit rate.

The test has been presented experimentally to experienced intelligence analysts. The results show that most probability assessors are able to improve their average score for assessments as a result of taking the test. Theoretically, the goal in the test should be to achieve a minimun error score. Users should keep that general goal in mind but depart from it to the degree that it is necessary in order to try out differing strategies of risky versus cautious attitudes toward probability assessment. The proper goal is to eliminate biases in assessing probabilities.

3.0 TECHNICAL OPERATIONS

This section explains in detail how a user interfaces with the SCORE software.

3.1 Options Available Prior to the Test

Once the program is loaded, the system immediately asks whether the user desires trial-by-trial feedback. That feedback consists of informing the user whether the answer was correct and the number of points gained or lost.

The system will then list the various question sets available, their subject matter (e.g., general, almanac, sports, etc.), their level of difficulty, and the number of questions comprising each set. The user will then be asked to designate the desired question set. The user may begin and end with any specific questions in the chosen question set. That is done by specifying the numbers of the first and last questions to be used.

The user is then given the opportunity to view instructions on how to take the test. Those instructions explain how to answer a question, how to edit an answer if the user has made a mistake in entering it, and a list of the commands available while taking the test. The instructions differ slightly depending on whether or not trial-by-trial feedback has been requested.

Questions are answered as follows. The user must decide which of the two possible answers is more likely to be correct and then type the number of that question. Next, the user types a space and then an assessment as to the

probability that the specified answer is indeed correct. The probability must be between .5 and 1, inclusive. Once the user has typed these values, the carriage should be returned so the answer can be recorded and the next question asked. If trial-by-trial feedback is being given, the system will inform the user of the number of points to be won or lost on the question, and will give the user the opportunity to adjust the response before proceeding.

Three commands are available to the user while taking the test. These are H (Help), F (Feedback), and S (Stop). "H" simply displays the available commands and an explanation of their functions. "F" causes the results up to that point to be analyzed and displayed. The results that are displayed are explained in Section 3.3. "S" allows the user to end the session prematurely.

Once the instructions have been displayed, the user may return the carriage to begin the test.

3.2 Taking the Test

When the first question appears, the user must type in the answer and associated probability, separated by a space, before returning the carriage.

If the user has previously requested no trial-by-trial feedback, the system will simply present the next question. If trial-by-trial feedback has been requested, the system will display (1) the number of points the user will win if the answer is correct and (2) the number of points the user will lose if the answer is incorrect. If the user does not like the odds inherent in the points-won-versus-points-lost comparison, there is an opportunity to change the answer. Typing a new answer and probability assessment will cause

the system to display the new number of points to be won or lost. If instead the user returns the carriage without typing anything, the system will display (1) the correct answer and (2) the net result in terms of user gain or loss. Returning the carriage will then cause the next question to be displayed.

Instead of answering any specific question, the user may instead use any of the three commands, "H," "F," or "S," as discussed in Section 3.1. The available feedback is of the same form as that received when the test is over, as explained in the next section.

3.3 The Test Results

Once a testing session is completed, the computer will display a graph exhibiting the results of that session. The user's responses are separated into ten categories: .5 to .54, .55 to .59, etc. Ideally, the user would correctly answer 52% of the responses in the .5 to .54 group correct, 57% in the .55 to .59 group, and so on. The graph displays this ideal result as well as the user's actual result.

The computer also displays how well the user has done compared to a perfectly calibrated analyst. This is stated as a percentage of the optimum, with the general knowledge component of the score removed. That is, this score only measures the user's calibration, or ability to use the probability scale correctly.

When the user returns the carriage, the system will state that the program is terminated. Instructions for restarting the program are displayed.

4.0 USE OF THE SCORE PROGRAM

This section demonstrates the use of SCORE. The first example assumes that no trial-by-trial feedback is requested. The second example involves trial-by-trial feedback.

4.1 Example Without Trial-by-Trial Feedback

The figures used in this section are representations of possible input and output formats. Other display formats would be equally suitable. In all figures, user inputs have been underlined for the purpose of clarity.

First, the user must load the program. The system will immediately ask whether trial-by-trial feedback is desired, as shown in Figure 4-1.

Do you want trial-by trial feedback during program operation?
Enter "Yes" or "No": *No*

Figure 4-1
SELECTING NO TRIAL-BY-TRIAL FEEDBACK

The system will then inform the user about the available question sets and request the user to designate one. This process is shown in Figure 4-2.

At present, sets of questions are available as follows:

SET NO.	SUBJECT	DIFFICULTY LEVEL	NO. QUESTIONS
1	ALMANAC	MODERATE	100
2	ALMANAC	MOD. DIFFICULT	100
3	GENERAL	MODERATE	100

ENTER THE DESIRED SET NO.: 1

Figure 4-2
SELECTING A QUESTION SET

The user is then permitted to use only a portion of the set, if so desired. Figure 4-3 illustrates this option.

The full set of 100 questions has been read into memory. If you wish, you may select any portion of the complete set.

ENTER NO. OF FIRST QUESTION TO BE USED: 1
ENTER NO. OF LAST QUESTION TO BE USED: 100

Figure 4-3
DEFINING PORTION OF QUESTION SET TO BE USED

Next, the system allows the user to obtain instructions on use of the program. This option appears in Figure 4-4.

Do you need instructions? Enter "Yes" or "No": Yes You will be asked a number of questions with two possible answers shown for each. For each question you will need to enter from the keyboard the following data:

(1) THE NUMBER OF THE ANSWER WHICH YOU FEEL IS MORE LIKELY TO BE CORRECT,
(2) A SPACE,
(3) THE PROBABILITY (a number in the range 0.5 to 1.0) THAT YOUR CHOICE IS CORRECT, AND
(4) THE "CARRIAGE RETURN" COMMAND.

Three commands are usually available while using this program:

'FEEDBACK' WILL REQUEST PERFORMANCE INFORMATION.
'STOP' WILL TERMINATE THE PROGRAM.
'HELP' WILL LIST THE COMMANDS.

Only the first letter of any command need be entered.

(Return carriage to continue)

Figure 4-4
THE INSTRUCTIONS

The system will next ask the first question. The user must respond, as explained in the instructions. At any point, the user may request performance information or a list of commands, or may terminate the program, instead of answering a question. Figure 4-5 depicts the user responses to the first few questions.

Which President was known as "Old Rough and Ready?"

1) ZACHARY TAYLOR
2) ANDREW JACKSON

2 .6

Which is the larger country? (In terms of area)

1) FRANCE
2) SPAIN

1 .85

Buddism had its origins in

1) CHINA
2) INDIA

2 .95

The ancient Mayan empire was located in

1) PERU
2) MEXICO

2 1.0

"Probity" means

1) WEALTH
2) INTEGRITY

2 .75

Figure 4-5
ANSWERING QUESTIONS

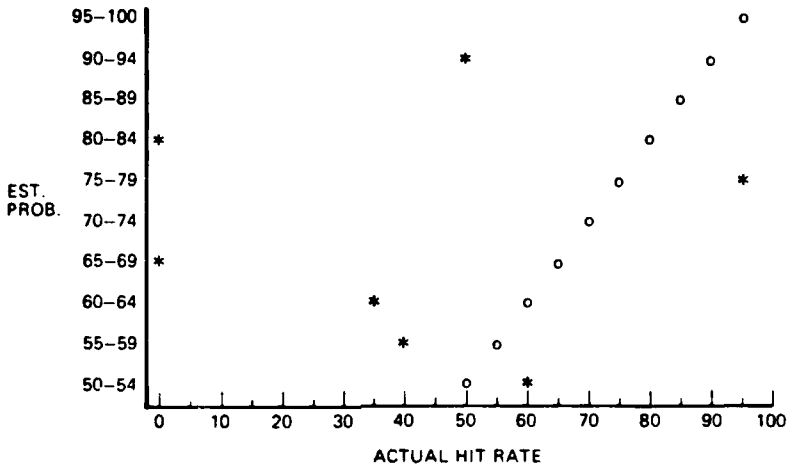
After twenty or so questions, the user might decide to get some feedback. This is done by typing an "F," as depicted in Figure 4-6.

Which President served the fewest days in office?

1) **WILLIAM H. HARRISON**
2) **JAMES GARFIELD**

F

The next graph will indicate how well observed hit rates agree with your assessments . . .



O OPTIMAL PERFORMANCE, *. ACTUAL PERFORMANCE

Your overall performance has been measured during the test by an average score. This score, however, results from two factors:

(1) **YOUR KNOWLEDGE OF THE SUBJECT AREA, AND**
(2) **YOUR ABILITY TO USE THE PROBABILITY SCALE CORRECTLY.**

Since the second factor is the one of interest here, your knowledge of the subject area can be "factored out" and a resulting score can be assigned that represents how near you came to the score you could have obtained with perfect use of the probability scale.

**** You obtained a score equal to 76.8 percent of your maximum obtainable score.*

Press return carriage to proceed.

Which President served the fewest days in office?

1) **WILLIAM H. HARRISON**
2) **JAMES GARFIELD**

1 1

Figure 4-6
REQUESTING FEEDBACK

Note that the system returns to the previous question after supplying the feedback information.

Finally, after the 100th question has been answered, the system will display the results of the test. Figure 4-7 depicts this process.

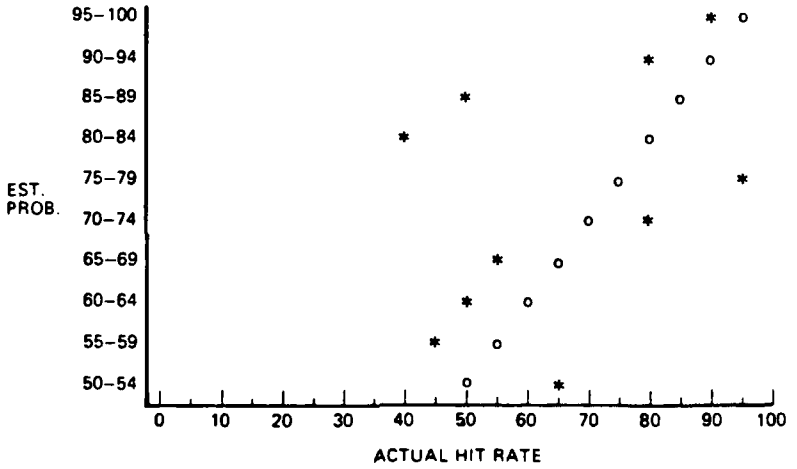
The painter, Peter Paul Reubens, was

1) DUTCH
2) FLEMISH

2 .68

Set of questions completed.

The next graph will indicate how well observed hit rates agree with your assessments . . .



Your overall performance has been measured during the test by an average score. This score, however, results from two factors:

(1) YOUR KNOWLEDGE OF THE SUBJECT AREA, AND
(2) YOUR ABILITY TO USE THE PROBABILITY SCALE CORRECTLY.

Since the second factor is the one of interest here, your knowledge of the subject area can be "factored out" and a resulting score can be assigned that represents how near you came to the score you could have obtained with perfect use of the probability scale.

**** You obtained a score equal to 78.8 percent of your maximum obtainable score.*

Press return carriage to proceed.

Figure 4-7
FINAL FEEDBACK

Note that the operator's actual performance was reasonably close to the ideal. The second number of responses contributing to the interim results prevented these interim results from appearing at all close to the optimal. After all, if only two responses are received in the 75-79% grouping, either 0, 50, or 100% (nowhere near 75%) of the responses must be correct.

Finally, the program will terminate and inform the user how to start again, as shown in Figure 4-8.

Program terminated, to restart, enter "Start."

START

Figure 4-8
TERMINATING AND RESTARTING THE PROGRAM

Here, the user restarts the program to do an example involving trial-by-trial feedback.

4.2 Example Including Trial-by-Trial Feedback

After restarting the program, the operator is again asked whether trial-by-trial feedback is desired. After responding positively, the user designates the question set and specific questions to be used. This sequence is depicted in Figure 4-9.

OPERATOR: Do you want trial-by-trial feedback during program operation; Enter "Yes" or "No": Yes

At present, sets of questions are available as follows:

SET NO.	SUBJECT	DIFFICULTY LEVEL	NO. QUESTIONS
1	ALMANAC	MODERATE	100
2	ALMANAC	MOD. DIFFICULT	100
3	GENERAL	MODERATE	100

ENTER THE DESIRED SET NO.: 3

The full set of 100 questions has been read into memory. If you wish, you may select any portion of the complete set.

ENTER NO. OF FIRST QUESTION TO BE USED: 1
ENTER NO. OF LAST QUESTION TO BE USED: 25

Do you need instructions? Enter "Yes" or "No": Yes

You will be asked a number of questions with two possible answers shown for each. For each question you will need to enter from the keyboard the following data:

(1) THE NUMBER OF THE ANSWER WHICH YOU FEEL IS MORE LIKELY TO BE CHOSEN
(2) A SPACE,
(3) THE PROBABILITY (a number in the range 0.5 to 1.0) THAT YOUR CHOICE IS CHOSEN
(4) THE "CARRIAGE RETURN" COMMAND.

The computer will respond by telling you how many points you will win if your answer is correct and how many points you will lose if your answer is incorrect. You may now either type in a new answer and probability (if you are unhappy with your possible gain or loss), or you may return the carriage without typing anything to officially record your response.

Editing is possible prior to pressing the 'Execute' key by

(1) USING THE DARK KEYS IN THE TOP ROW LABELED '←' AND '→' TO POSITION THE CURSOR (the flashing position indicator) AND,
(2) KEYING IN NEW CHARACTERS TO REPLACE THE OLD.

Three commands are usually available while using this program:

'FEEDBACK' WILL REQUEST PERFORMANCE INFORMATION.
'STOP' WILL TERMINATE THE PROGRAM.
'HELP' WILL LIST THE COMMANDS.

Only the first letter of any command need be entered.
(Return carriage to continue)

Figure 4-9

STARTING THE TEST

Again, the system will begin asking questions. After the user responds, the system will give feedback concerning the correct answer. Figure 4-10 depicts this process for the first few questions.

Which is taller:

1) ST. PAUL'S CATHEDRAL
2) EIFFEL TOWER

1 .6
Win: 9.0 or Lose: 11.0
Correct Answer: 2 – you lose 11.0 points.
Return carriage for next question.

At the start of World War I, Romania

1) DECLARED ITS NEUTRALITY
2) JOINED THE ALLIES

2 .8
Win: 21.0 or Lose: 49.0
Correct Answer: 1 – you lose 49.0 points.
Return carriage for next question.

The Renaissance may be described as a period characterized by

1) A PASSIONATE AIMING AT THE HELLENIC IDEAL
2) INDIVIDUAL SELF-RESTRAINT

1 1
Win: 25.0 or Lose: 75.0
Correct Answer: 1 – you win 25.0 points.
Return carriage for next question.

The most important factor in England's rise to power in the 16th Century was the

1) DESTRUCTION OF THE SPANISH NAVAL POWER
2) RICHES ACQUIRED THROUGH EXPEDITIONS TO THE NEW WORLD

1 1
Win: 25.0 or Lose: 75.0
Correct Answer: 1 – you win 25.0 points.
Return carriage for next question.

Figure 4-10
RESPONDING AND OBTAINING FEEDBACK

Figure 4-11 shows the process of seeking help and stopping the test.

The Jacquerie was a

1) **PEASANT UPRISING**
2) **SOCIETY IN PARIS**

H
"Help" command is to be used to list available commands. Commands are: "Help," "Feedback," and "Stop." Only the first letter need be entered.
Please try question again:

The Jacquerie was a

1) **PEASANT UPRISING**
2) **SOCIETY IN PARIS**

S
Do you want feedback before stopping? Enter "Yes" or "No": No
Program terminated. To restart, enter "Start."

Figure 4-11
STOPPING THE TEST

Had the user continued the test, a final graph and score would have been provided, as in the previous example.

5.0 ABRIDGED USERS MANUAL

This section is designed for the user who is already familiar with SCORE. It describes the essential elements of SCORE and explains how to use the program.

5.1 The Purpose of SCORE

SCORE is a program designed to test probability assessment skills and to train assessors to assess probabilities more accurately. It accomplishes this by providing feedback on how well a user assesses probabilities and by pinpointing specific weaknesses.

The program consists of a series of questions accompanied by two answers, one of which is correct. The user must identify which of the two answers is correct and specify the degree of certainty in the form of a probability. After a series of questions and responses, the program will display a graph showing the user's calibration error, followed by an overall performance measure stated as a percentage of optimal performance.

5.2 Using the Program

Once the program is loaded into the computer, the user must specify whether trial-by-trial feedback is desired. One of the available question sets must be selected, and the desired number of questions from that set specified. Directions for using the system may be requested, if desired.

Once the program begins, the user responds to the questions by typing the number of the answer believed to be correct and the probability that it actually is correct.

Probabilities range from .5 to 1. Returning the carriage will, in the case of trial-by-trial feedback, cause the system to inform the user how much he or she stands to win or lose and will give the user the opportunity to change the original response. When this is done, the system will reveal whether or not the user's answer was correct and how many points were won or lost. The system will then proceed to the next question. In the case of no trial-by-trial feedback, the system will ask the next question immediately after the answer to the previous question is specified.

At any time during the session, the user may request a list of commands, a synopsis of the results up to that point, or end the test by typing "H," "F," or "S," respectively, instead of answering a question. When the test is finished, the test results will be displayed. The user may then choose to end or restart the program.